DOCUMENT RESUME

ED 422 363 TM 028 920

AUTHOR	Skaggs, Gary; Bourque, Mary Lyn
TITLE	Overview of the Most Difficult Technical Issues on the VNT.
INSTITUTION	National Assessment Governing Board, Washington, DC.
PUB DATE	1998-04-00
NOTE	12p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 12-16, 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Achievement; Achievement Tests; Classification; Difficulty Level; Disabilities; *National Competency Tests; Political Influences; *Psychometrics; *Sampling; State Programs; *Test Reliability; *Testing Problems; Testing Programs
IDENTIFIERS	*Linking Metrics; National Assessment of Educational Progress; *Voluntary National Test

ABSTRACT

Political and legislative pressures have posed a number of measurement issues and challenges to the development of sound, valid voluntary national tests (VNTs). This paper focuses on what appear to be the most difficult technical issues related to the VNT proposed by President Clinton in 1997. Technical issues refer to psychometric issues, as opposed to administrative or policy issues. The requirement that the VNT be linked to the National Assessment of Educational Progress (NAEP) to the maximum extent possible, in addition to linking with the Third International Mathematics and Sciences Study (TIMSS), poses a number of technical issues. These include: (1) the lack of experience with linking reading test results; (2) the type of linking methodology; (3) linking design; (4) model specification; (5) the stability of linking; and (6) the invariance of linking. The desired result of the VNT is a classification of students according to NAEP achievement levels. Complexities of the classification process include the reliability of classification and test length and selection of items. Conducting pilot and field testing also poses a number of issues, including the problems of drawing nationally representative samples in years in which pilot, field, and operational tests will all occur. Another set of problems relates to inclusion of students with disabilities and accommodations some groups of students, including those of limited English proficiency, may require. How to aggregate data and how to ensure that the VNT has no adverse impact on NAEP and state and local testing are problems that must be resolved. Other technical issues may come to light as the development process moves forward. (Contains 11 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 422 363

Overview of the Most Difficult Technical Issues on the VNT

Gary Skaggs and Mary Lyn Bourque
National Assessment Governing Board

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Gary Skaggs
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM028920

Paper presented at the 1998 Annual Meeting of the National Council on Measurement in Education, San Diego.

Overview of the Most Difficult Technical Issues
on the Voluntary National Test

Gary Skaggs and Mary Lyn Bourque
National Assessment Governing Board

In President Clinton's 1997 State-of-the-Union address, he proposed the development of voluntary national tests to measure the achievement of fourth grade reading and eighth grade mathematics. According to the Department of Education, the purpose of these tests is to assist public education in raising academic standards nationwide. The Department of Education, through the National Center for Education Statistics (NCES), proceeded with a competitive bid process and contracted with the American Institutes for Research (AIR) for developing the tests.

However, in November 1997, Congress passed P.L. 105-78. This law stated that the new Voluntary National Tests (VNTs) shall:

- be based on the same content and performance standards as are used for the National Assessment of Educational Progress (NAEP)
- be linked to NAEP to the maximum extent possible (note that the 8th grade VNT-M will also be linked to TIMSS)
- differ substantively from NAEP only when the individual nature of the VNT requires it.

The law also gave the National Assessment Governing Board (NAGB) exclusive authority over test development, including both setting policy on the VNTs and managing the contract for their development. Finally, and critically, the law prohibited the use of 1998 fiscal year funds for any pilot or field testing.

In January 1998, NAGB decided to modify the contract with AIR in certain ways, most importantly to revise the test development timeline so that pilot testing, field testing, and operational testing would be conducted in March of 1999, 2000, and 2001, respectively. The end product will be a 90-minute test with two 45-minute sessions. An examinee who takes the test will obtain a score that will predict which of the NAGB achievement level categories (below basic, basic, proficient, advanced) the examinee would obtain had he/she taken NAEP. For 8th graders, there will also be a predicted score for TIMSS. The scoring will then place an examinee according to fairly demanding national standards and in the case of TIMSS international standards. The examinee will therefore know how he/she compares to the state, nation, and other nations in relation to a high set of standards.

The above decisions have posed a number of measurement issues and challenges to the development of sound, valid tests. As part of a symposium that addresses these issues, this paper will focus on describing a number of what appear at the present time to be the most difficult technical issues related to the VNT. Technical issues here will refer to psychometric issues as opposed to administrative or policy issues.

How to Link the VNT to NAEP and TIMSS

The requirement that the VNT shall be linked with NAEP to the maximum extent possible, in addition to linking with TIMSS, poses a number of technical issues. Some research has been done attempting to link various assessments to NAEP. Table 1 below lists the most widely publicized of these research studies.

Table 1
Research Studies Attempting to Link Assessments to NAEP

Authors	Test Linked to NAEP	Data Design	Linking Method	Results
Beaton & Gonzalez (1993)	IAEP	Trial State Assessment – 1990; IAEP – 1991. Separate samples; Math; 13 yr. olds	Statistical moderation using linear equating.	Obtained percents of populations of states and nations at NAGB achievement levels
Pashley & Phillips (1993)	IAEP	1992 sample of US 8th graders adm both IAEP and NAEP; Math	Projection	Obtained percents of populations of states and nations at NAGB achievement levels; Distribution of predicted NAEP scores was narrower than Beaton & Gonzalez, affecting percents above cutpoints
Ercikan (1993)	CAT/5; CTBS/4	1990 Trial State & 1990 CAT/CTBS; separate samples; 8th grade Math	Statistical moderation using equipercentile equating	Invariance was lacking. Projection functions from each state were significantly different from each other. There were also differences between predicted and actual distributions of NAEP scores. Results were attributed to differences between the two instruments, lack of motivation for taking NAEP, representativeness of NAEP sample, and differences in reliability.
Linn & Kiplinger (1993)	Four state tests	1990 Trial State Assessment, 8th grade Math; statewide tests; separate samples	Statistical moderation using equipercentile equating	Large differences between separate gender-based linkings. Linkings were also unstable across time. Invariance was thought to be due to content differences between state tests and NAEP.
Bloxom et al (1995)	ASVAB	Sample of military applicants adm both ASVAB and NAEP; 12th grade Math	Projection:	Invariance was lacking. Separate analyses by subgroups produced different projection functions. Also, motivation on NAEP was suspected to be low.
Williams et al (1995)	NC End-of-Grade	Feb. 1994 sample of 8th graders adm both NC EOG & NAEP; Math	Projection	Attempted to calibrate, but relationship between NC test and NAEP was not strong enough to justify (r=mid.80s). Both instruments were based on NCTM standards but different frameworks. Results also showed a lack of invariance for ethnic subgroups. Observed and predicted NAEP distributions were very similar. Possible motivational differences and the stability of the linking over time are mentioned as challenges to be explored further.

This table illustrates a number of difficult technical issues related to the proposed linking of the VNTs to NAEP and TIMSS.

Little Experience with Linking Reading. All of the above studies involve linking tests to NAEP in mathematics. To our knowledge, no published study has attempted to link to NAEP in reading, although some preliminary work was done in Kentucky and a more extensive study is planned in Louisiana. In order to provide a broad range of passages on the VNT, the VNT will contain shorter reading passages than typically found in NAEP. Furthermore, the use of intertextual items has been proposed. These items have been previously used only at 8th and 12th grades.

Type of Linking Methodology. Using the framework of Linn (1993) and Mislevy (1992), the strongest possible link between the VNT and NAEP or TIMSS would be an equating function in which each possible score on the VNT would be equivalent to a score on the NAEP or TIMSS scale. Equivalent in this case means that scores on each assessment would have the same meaning with the same degree of precision, or in other words, it would be a matter of indifference to an examinee whether he/she would take the VNT or NAEP/TIMSS.

Since the VNT will differ from NAEP and TIMSS in some important ways, notably in test length, equating the VNT to NAEP or TIMSS will not be possible. The major technical issue is: what alternative linking methodology can be used and will the linked estimate satisfy the needs of policy-makers?

The next strongest type of statistical link between two instruments is calibration. Calibration requires that the two assessments measure essentially the same construct, but the precision of measurement need not be the same. It is doubtful that the VNT/TIMSS link will qualify as calibration. This is because the two assessments differ in content and in a number of administrative procedures, for example, in the use of formulas and calculators. It is likely then that the math VNT will measure related but decidedly different constructs from the TIMSS.

For NAEP, however, there is some hope that calibration can be accomplished. The specifications for the VNTs are intended to be as similar to NAEP as possible. The two assessments are based on the same NAEP frameworks and employ very similar administrative procedures. However, important differences remain. For both tests, the individual nature of the VNTs will necessitate a number of test form differences between it and NAEP. For reading, the VNT will contain shorter passages than are normally found in NAEP, and intertextual items are proposed. For mathematics, gridded items and machine-scorable drawing items are proposed.

The next strongest link is referred to as projection. If one has a data collection design in which examinees have taken both assessments (and this is the design proposed by ETS), then projection may be possible. Projection relaxes the requirements of calibration in that the two assessments need not measure the same construct, only constructs which are strongly related. Regression is the typical methodology for projection. Among the studies in Table 1 which used projection, correlations in the neighborhood of the mid-.80s were typically found.

The choice between calibration and projection will be made mostly on empirical grounds. However, the two methods will produce quite different results. First, projection is not symmetrical. Unlike calibration, using projection, a score on the VNT will predict a NAEP scale score, but that NAEP scale score will not in turn predict the original score. Second, if regression is used for projection, predicted scores will regress to the mean of the NAEP scale score distribution. High scores on the VNT will predict not quite so high scores on NAEP, and low scores on the VNT will predict higher scores on NAEP. This has implications for both data aggregation and for studying subgroups who are above or below the mean. Using projection, it will be more difficult for examinees to be predicted to be advanced. If the relationship between the VNT and NAEP is weak enough, then it may be impossible for anyone to be advanced, even with a perfect score on the VNT! While psychometricians are well-aware of these issues, it is not clear that the general public is.

Linking Design. The current plan is to obtain scores on both the VNT and NAEP on a subsample of the field-test sample, in March 2000. For a separate subsample of the same field test, scores will be obtained for both the VNT and TIMSS. From a psychometric perspective, an ideal linking design would have these assessments administered in a counterbalanced order and within a very short time of each other. This is not possible with the VNT. NAEP will be administered in February, the VNT will be administered in March, and TIMSS will be administered in April.

There will conceivably be some content covered by both the VNT and NAEP for which students will receive instruction between the two testing times. The same holds true for the VNT and TIMSS. If this is true, then the strength of the linking is weakened. Some norm-referenced testing programs provide adjusted norms for each week that test administration time is removed from the time the norming sample was tested. At this time, we do not have a way of adjusting for instructional differences either through the linking design or linking methodology.

Model specification. The VNT will be based on unidimensional IRT models while NAEP is based on multidimensional models. The NAEP Mathematics Assessment is based on five independently scaled content strands. The NAEP Reading Assessment is based on three independently scaled reading situations, two of which are tested at the 4th grade. Because of test length, each VNT will be scaled as a single unidimensional construct, even though it will contain items from all of the independently scaled NAEP subscales.

Another aspect of this issue is that the NAEP scale is designed to span 4th through 12th grades. To accomplish this vertical scaling, examinees are presented with items that are also presented to the adjacent testing level. Thus, 4th graders are administered items that are also administered to 8th graders: Eighth graders are likewise administered items also administered to 4th and 12th graders. The VNT test specifications require that all test items are grade appropriate. This means that the difficulty level and to some extent the content of the items taken by VNT examinees and NAEP examinees at a given grade level will be different. It also means that, if calibration is attempted, the VNT scaling will map onto only a portion of full NAEP scale with some potential for basal or ceiling effects within the VNT forms.

Stability of linking. The current plan is to use data from the first field test, in March 2000, to conduct the linking study. However, the following year, the VNT will be administered operationally for the first time. It is likely that the relationship between the VNT and NAEP (and TIMSS) will change over time. This could be due to several factors including familiarity with the VNT, greater motivation on the VNT than NAEP, greater test preparation for the VNT, and adjustment of local curricula to align with the VNT. ETS has proposed to reconduct linking studies as a part of the 2001 and 2002 field tests. However, if the linking does change, does one use the new linking? If so, how will that affect any longitudinal comparisons?

Invariance of linking. Several of the studies listed in Table 1 reported a problem with invariance. In other words, if the linking analysis is repeated for different subgroups, such as for gender and race/ethnicity, different linking functions result. This is problematic because if a single linking is used for all examinees, then the test could appear to be biased for or against certain subgroups. This problem may even extend to having different linking functions for different states.

To make matters worse, in producing plausible values, NAEP uses these subgroup designations as conditioning variables. This makes a lack of linking invariance even more likely for any assessment to be linked to NAEP. To separate true test bias from invariance due to the conditioning process is a serious psychometric challenge.

How to classify individuals into achievement levels

The desired end product of the VNT is a classification of students according to the NAGB achievement levels used in NAEP. Given resolution of the linking issues discussed above, the calculation of the linking function is fairly straightforward statistically. However, there are potential complexities in the classification process.

Reliability of classification. The classification of individuals into achievement levels will have a certain degree of uncertainty due to issues noted above concerning the linking. Hopefully, the degree of uncertainty will be acceptable enough for most examinees. How to report the level of uncertainty to the public in an understandable way is a significant challenge. ETS has proposed two possible mechanisms for reporting: a predicted NAEP score range and a probability of being in each of the achievement level categories. The challenge is to convey information about the accuracy of the predicted achievement level without seeming to lack confidence in the prediction itself.

Test length and selection of items. The intent of NAGB is to support strongly prediction of all three achievement levels. However, the proposed VNT will be a relatively short test for accurate classification according to three cutpoints. This will require a very careful selection of items into test forms. The items must span a wide range of proficiency. The Basic and Proficient cutpoints are located in populous parts of the NAEP distribution, but the Advanced cutpoint is fairly extreme (Only about 2 percent of 4^{th} and 8^{th} graders are Advanced.). On NAEP, a great deal of the ability to classify at the advanced level is due to items also administered to the next highest

testable grade, that is, 8th grade reading items administered to 4th graders and 12th grade math items administered to 8th graders. Only grade-level appropriate items will be featured on the VNT. It may therefore be difficult to develop enough items to measure the advanced level accurately.

How to Conduct Pilot and Field Testing

The current plans for the VNTs call for pilot-testing in March 1999, field-testing in March 2000, and operational testing in March 2001. Currently, a second testing cycle will consist of pilot-testing in March 2000, field-testing in March 2001, and operational testing in March 2002. A third testing cycle will consist of pilot-, field-, and operational testing in March of 2001, 2002, and 2003, respectively. The purpose of the pilot of test is to tryout test items. The purpose of the field test is to tryout test forms, testing conditions, and scoring mechanisms under conditions as realistic as possible. The field test will also provide data for equating alternate forms and linking the VNT to NAEP and TIMSS.

Extensive plans have been developed for both the pilot and field tests, and tens of thousands of students will involved in these tests. A potentially very difficult problem can occur in the years 2001 to 2003. In those years, a pilot test, field test, and operational test will all be administered in March of those years. Current plans for the field tests include drawing nationally representative samples. However, only students not taking the operational VNT can be targeted for the field test. The operational test will not be nationally representative because the VNT will be voluntary. If participation in the operational VNT is strong enough, it will be difficult to draw a nationally representative sample from those students who are left. It is quite likely that it will difficult to sample enough students from certain strata. For example, at the present time, many large cities have signed up for the VNT.

A related issue is that it may be difficult to enlist enough schools to participate in the pilot test. Under current plans, students participating in the field-test would receive some form of test results. For pilot testing, they would not. Schools would probably much rather participate in the field test.

How to Provide Inclusions and Accommodations and Maintain Psychometric Integrity

A major issue that will be debated extensively in the coming months is what policy to set on inclusions and accommodations. Inclusions and accommodations generally relate to two groups of people, those with disabilities and those with limited English proficiency.

Students with Disabilities. Decisions on who to include in the test population and who to accommodate have an impact on the linking of the VNT to NAEP and TIMMS. If the two assessments have different rules on inclusions and accommodations, then the linking will not be based on the same populations of examinees. The sample used for linking will be biased from the perspective of either or both examinee populations.

As a part of the 1996 NAEP Mathematics Assessment, the impact of alternate rules for inclusions and accommodations on participation rates was examined. Essentially, revised rules for inclusions had no significant impact on participation rates or proficiency estimation.

However, provisions for a number of accommodations along with the revised inclusion rules significantly improved participation rates in the grades that will be involved in the VNT – 4th and 8th grades. A report on how the NAEP items worked under new accommodation rules will be available soon. Some preliminary findings suggest that national aggregates are affected minimally by greater accommodations. On the other hand, we know little about the impact of the rules on individual students or on smaller aggregations of data, such as classes or schools.

An important question to ask is: will the VNT be measuring the same reading and math constructs for students receiving accommodations? A sticky aspect of this issue concerns a group of students who might have taken the VNT without accommodations but can take advantage of them if available. Which is the more valid score for them? Do they have an advantage over other students taking the test?

To some extent, requirements for greater inclusion and accommodations may conflict with linking to NAEP to the maximum extent possible. The decision on where to strike a balance between competing objectives may be largely a legal one, but it has major technical implications.

Limited English Proficiency (LEP). The discussion on this issue centers on whether the VNT should be offered in languages other than English, and there are some groups of people who feel very strongly that this should happen. While direct translations of the math VNT seem feasible, there are some issues regarding the reading VNT. First, among them is a validity issue. Is reading in another language the same construct as reading in English? One could picture, for example, that a literary passage directly translated into a second language could become easier or more difficult in the second language. Some have proposed to use, for each language, passages that occur directly in that language, thus by-passing translation. If this occurs, can these non-English forms be equated to the English forms?

There is some research to suggest that equating forms in different languages is not necessarily straightforward. Angoff & Cook (1988 attempted to equate the Scholastic Aptitude Test (SAT) to a Spanish version of the test called the Prueba de Aptitud Academica. The equating worked reasonably well for the quantitative scale, but serious problems occurred on the verbal scale.

There are numerous technical issues surrounding translations of tests. For example, some languages will require more words than English for direct translations of passages. Sireci et al. (1997) found evidence of this in a test translated from English to German. Text length will have an impact on testing time and speededness. Another issue surrounds the intended use of authentic passages on the VNT. It may be that some passages, written originally in English, will contain words or phrases which have no easy direct translation, or the translated words will have other meanings in the second language that do not exist in English (e.g., "once in a blue moon"). Additionally, examinees proficient in a second language may nevertheless be more familiar with certain English terms than those in their own language. This implies that a bilingual test be considered (i.e. questions printed in both languages). As Sireci (1997) has pointed out, the great research challenge in all of these multi-language issues is to separate differences between translated tests and differences from differences between the two language target populations.

If the VNT is given in languages other than English, there is also likely to be an issue of describing the achievement levels. Currently, for each NAEP assessment, considerable effort is given to developing descriptions for each achievement level – basic, proficient, and advanced.

They describe what students at each level at each grade level of the assessment (4, 8, and 12) know and can do. Many languages are structured very differently from English. Will the same achievement level descriptions for English apply to these other languages? If the descriptions are different, does it make sense to put these tests on the English forms scale? If not, how do we represent at least equivalent levels of achievement?

How to Aggregate Data

Although NAGB has not yet set a policy on aggregation of VNT scores, some schools and districts may want to aggregate data by classroom, school, and district. This will invite comparisons between the aggregate distribution and the distributions reported by NAEP for the states and nation and for TIMSS internationally. However, simple aggregation of VNT scores may produce estimates of population distributions with more or less variance than those produced by NAEP and TIMSS. For one thing, if regression is used as the projection methodology, predicted scores will be more compact around the mean. This will cause a tendency for the aggregate to underestimate the percent of advanced and below basic students.

Second, as Johnson and Mazzeo (1998) point out in their paper that is a part of this symposium, NAEP/TIMSS plausible values and NAEP/TIMSS scores predicted from the VNT embody different error structures. For one thing, the VNT for individuals will be a very reliable instrument while, for NAEP and TIMSS, measurement error is so great as to preclude getting any degree of individual reliability. The degree of reliability will influence the variance of observed test scores. If the VNT and NAEP/TIMSS differ in reliability at the individual level, as they almost certainly will, then the simple aggregation of observed VNT scores will be a biased estimate of the population distribution of NAEP/TIMSS scores and consequently an over- or underestimate of the percents of students above achievement level cutpoints.

A recommendation to users could be that conditioning should be used when aggregating VNT scores. The conditioning also needs to be carried out for each aggregation since the degree of bias varies on each intended aggregation (i.e. class, school, district). However, this is a highly complex procedure. Very few school personnel understand the conditioning process. How will local school districts be able to carry out this procedure?

How to Ensure that the VNT Has No Adverse Impact on NAEP and State and Local Testing

The VNT will not operate in a vacuum. There is some concern that, if participation in the VNT is high enough, it may be more difficult to obtain school participation in the state and/or national NAEP. Or, one can conceive even for a push for VNT data to replace NAEP data (for 4th grade reading and 8th grade mathematics). Why give two tests that measure the same thing? We note in passing that participation in state NAEP has declined recently. A major reason for this is that schools feel that they are involved in too much testing already, and that the burden is too great.

The introduction of the VNT will have an additional impact on state and local testing programs. Many states and local school districts will want to re-examine their entire testing programs in light of the VNT. How do commercial, state, and locally developed tests compare to the VNT,

and what should the testing program look like in 4th and 8th grades? One possible outcome, discussed in several states, is to try to link state assessments to NAEP/VNT achievement levels. There are significant technical challenges associated with that type of linking.

Summary

Given the daunting list of technical issues discussed in this paper, one may ask if it is even possible to develop a reliable and valid Voluntary National Test. We note that a number of issues will be decided on non-technical grounds. For example, decisions about inclusions and accommodations will be made largely on legal grounds in addition to broad input from the public. Other issues, such as type of linking methodology, will be decided on the basis of empirical evidence. And finally, on some issues, limits are imposed based on practical concerns. An example of this type is shorter reading passages.

There may be additional technical challenges that we do not yet know about as the development process moves forward. If the Voluntary National Test is approved by Congress, it will be developed in an open and very public manner, and it will probably be the most closely scrutinized test in history. To the benefit of the VNT, this means that all stakeholders and the public at large will be heard in its development to an extent never before attempted. Difficult issues such as linking, data aggregation, and inclusions/accommodations will be handled so that the public is well-informed.

References

Angoff, W.H. & Cook, L.L. (1988). *Equating the Scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (Report 88-2). New York: College Entrance Examination Board.

Beaton, A.E. & Gonzalez, E.J. (1993). *Comparing the NAEP Trial State Assessment Results with the IAEP International Results*. Paper prepared for the Panel on the Evaluation of the NAEP Trial State Assessment of the National Academy of Education.

Bloxom, B. Pashley, P.J., Nicewander, W.A., & Yan, D. (1995). Linking to a large-scale assessment: an empirical evaluation. *Journal of Educational and Behavioral Statistics, 20*, 1-26.

Ercikan, K. (1993). *Predicting NAEP*. Unpublished manuscript. Monterey, CA: CTB MacMillan/McGraw-Hill.

Linn, R.L. (1993). Linking results of different assessments. *Applied Measurement in Education, 6*, 83-102.

Linn, R.L. & Kiplinger, V.L. (1995). *Linking statewide tests to the National Assessment of Educational Progress: Stability of Results*. Unpublished manuscript. Center for Research on Evaluation, Standards, and Student Testing: University of Colorado at Boulder.

Mislevy, R.J. (1992). *Linking educational Assessments: Concepts, Issues, Methods, and Prospects*. Princeton, NJ: Educational Testing Service.

Pashley, P.J. & Phillips, G.W. (1993). *Toward World-Class Standards: A Research Study Linking International and National Assessments*. Princeton, NJ: Educational Testing Service.

Sireci, S.G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practices, 16*, 12-19.

Sireci, S.G., Foster, D.F., Olsen, J, & Robin, F. (1997). *Comparing Dual-Language Versions of an International Certification Exam*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Williams, V.S.L., Billeaud, K., Davis, L.A., Thissen, D., & Sanford, E. (1995). *Projecting to the NAEP Scale: Results from the North Carolina End-of-Grade Testing Program*. Research Triangle Park, NC: National Institute of Statistical Sciences.